Earth and Space Science

RESEARCH ARTICLE
10.1002/2017EA000350

Key Points:
- We completed OSSEs to study the impacts of satellite and dropsonde data on global-average and storm-specific forecasts
- Removing satellite data degrades global average forecasts; adding dropsondes over a large idealized region mitigates this degradation
- Targeted dropsondes usually improve forecasts of three winter storms compared to a loss of satellite data, but results are case dependent

Improving Winter Storm Forecasts With Observing System Simulation Experiments (OSSEs). Part 2: Evaluating a Satellite Gap With Idealized and Targeted Dropsondes

Jason M. English[1,2], **Andrew C. Kren**[3], **and Tanya R. Peevey**[1,2]

[1]Cooperative Institute for Research in Environmental Sciences at the NOAA/OAR/Earth System Research Laboratory/Global Systems Division, University of Colorado Boulder, Boulder, CO, USA, [2]Global Systems Division, NOAA Earth System Research Laboratory, Boulder, CO, USA, [3]Cooperative Institute for Marine and Atmospheric Studies, University of Miami (formerly CSU/CIRA/GSD), Miami, FL, USA

Abstract Numerous satellites utilized in numerical weather prediction are operating beyond their nominal lifetime, and their replacements are not yet operational. We investigate the impacts of a loss of U.S.-based microwave and infrared satellite data and the addition of dropsonde data on forecast skill by conducting Observing System Simulation Experiments with the European Centre for Medium-range Weather Forecasts T511 Nature Run and the National Center for Environmental Prediction Global Forecast System Model. Removing all U.S.-based microwave and infrared satellite data increases Global Forecast System analysis error, global forecast error, and forecast error during the first 36 hr of three winter storms that impact the United States. Data from Suomi National Polar-orbiting Partnership contributes roughly one third of the total satellite impacts. Assimilating "idealized" dropsondes (sampling over a large region of the Pacific/Arctic Oceans) significantly improves global forecasts and forecasts for all three storms. Assimilating targeted dropsonde flight paths using the Ensemble Transform Sensitivity method for 15 verification dates/locations for the three storms improves roughly 80% of forecasts relative to the control and 50% of forecasts relative to their corresponding experiments without dropsondes. However, removing satellite data degrades only 30% of targeted domain forecasts relative to the control. These results suggest that targeted dropsondes cannot compensate for a gap in satellite data regarding global average forecasts but may be able to compensate for specific targeted storms. However, as with any study of specific weather events, results are variable and more cases are needed to conclude whether targeted observations—as well as satellite data—can be expected to improve forecasts of specific weather events.

Correspondence to:
J. M. English,
jason.english@noaa.gov

Citation:
English, J. M., Kren, A. C., & Peevey, T. R. (2018). Improving winter storm forecasts with Observing System Simulation Experiments (OSSEs). Part 2: Evaluating a satellite gap with idealized and targeted dropsondes. *Earth and Space Science*, *5*, 176–196. https://doi.org/10.1002/2017EA000350-T

Received 30 NOV 2017
Accepted 4 APR 2018
Accepted article online 23 APR 2018
Published online 11 MAY 2018

1. Introduction

The importance of satellite data to improve numerical weather prediction (NWP) medium-range forecast skill through reduction of initial condition errors is well known (e.g., Baker et al., 2005; Simmons & Hollingsworth, 2002). Unfortunately, numerous satellites are near or past the end of their nominal lifetime with either no replacement plan or a gap before their replacement is launched (Table 1). For example, the Suomi National Polar-orbiting Partnership (NPP) satellite mission's life expired in October 2016, and its replacement, the first Joint Polar Satellite System satellite, was not launched until November 2017, a 13-month gap. However, there are presently many satellites that are assimilated into a modern data assimilation system with some redundancy. Suomi-NPP measures microwave (MW) radiances from the Advanced Technology Microwave Sensor (ATMS) and infrared (IR) radiances from the Cross-track Infrared Sounder (CrIS). Suomi-NPP was placed in early afternoon orbit, providing similar data as the existing polar orbiting satellites with MW sounders (AMSU-A and MHS on NOAA 18/19, AMSU-A on Aqua, and AMSU-A on NOAA 15).

Numerous Observing System Experiments (OSEs) have been conducted to understand the impacts of satellite data on forecast accuracy. Several studies have found Northern Hemisphere forecast degradations in the National Center for Environmental Prediction (NCEP) models when various satellite data are removed: Removing ATMS (Zou et al., 2013), removing MW, IR, and radio occultation (McNally, 2012), and removing secondary MW and IR instruments (Boukabara et al., 2016) all resulted in forecast degradation. Other studies have found no significant impacts on Northern Hemisphere forecast skill: removing ATMS (Garrett, 2013) and removing U.S.-based MW instruments (Cucurull & Anthes, 2015). Multiple studies found satellite

Table 1
List of Satellite Instruments, Launch Dates, and Nominal Lifetimes

Satellite	Microwave	Infrared	Radio occultation	Launch date/nominal lifetime (years)	Replacement satellite	
					Name	Launch date/nominal lifetime (years)
Aqua	AMSU-A	AIRS		May 2002/6	NA	NA
MetOp-A	AMSU-A, MHS	HIRS4, IASI	GRAS	Oct 2006/5	MetOp-B	Sep 2012/6
Suomi-NPP	ATMS	CrIS		Oct 2011/5	JPSS-1	2017/7
NOAA 15	AMSU-A			May 1998/2	NOAA 16	Sep 2000/2
NOAA 18	AMSU-A, MHS			May 2005/2	NOAA 19	Feb 2009/2
NOAA 19	AMUS-A, MHS	HIRS4		Feb 2009/2	NOAA 20	2017/7
Metosat-9		SEVIRI		Dec 2005/7	Metosat-10	Jul 2012/5
GOES-13		SNDR D1-D4		May 2006/10	GOES-16	Nov 2016/15
DMSP F16	SSMIS			Oct 2003/14	DMSP F17	Nov 2006/11
GRACE-A			JPL Blackjack	Mar 2002/5	GRACE-FO	2018/5
COSMIC			JPL Blackjack	Apr 2006/5	COSMIC-2	2017 and 2020/5
TerraSar-X			JPL Blackjack	Jun 2003/5	TerraSar-NG	2018/5
C/NOFS			CORISS	Apr 2008/2	NA	NA

Note. Satellites in this table with a launch date prior to 2015 are assimilated into the Global Forecast System model.

observations to be more important in the Southern Hemisphere than the Northern Hemisphere (Baker et al., 2005; Cucurull & Anthes, 2015; Lord et al., 2016; McNally, 2012). Given the large number of satellites that are near or past their nominal lifetime, there remains a risk to degradation of forecasts, but it remains unclear how large the impacts will be at various locations.

In theory, dropsonde data might be able to help mitigate the impacts of a gap in satellite data on forecasts, although this has not yet been studied and may require an unrealistic number of dropsondes. In practice, dropsondes are typically part of a "targeted observations" campaign to improve forecasts of specific weather events. The goal of a targeted observation effort is to identify regions where a reduction in analysis error can improve forecasts over a specified verification region. OSEs with targeted observations have been employed for hurricanes and tropical storms (Aberson, 2008, 2010), polar storms (Irvine et al., 2009, 2011), and extratropical storms (Langland, Gelaro, et al., 1999; Langland, Toth, et al., 1999; Szunyogh et al., 2000, 2002). These studies generally find small improvements to forecasts on average, but results are case dependent and sometimes forecasts are neutral or degraded. Several reviews of targeted observations have been completed and generally conclude that a small majority of targeted observations improve forecasts (Gelaro et al., 2010; Langland, Toth, et al., 1999; Lorenc & Marriott, 2014; Majumdar, 2016). However, results from the 2011 Winter Storm Reconnaissance program found generally neutral results (Hamill et al., 2013). While the differences between "small majority" and "neutral" may be mathematically small, it can have significant impacts on whether targeted observation campaigns may be deemed worthwhile.

Observing System Simulation Experiments (OSSEs) can be conducted in addition to or instead of OSEs. OSSEs use *simulated* observing systems where a forecast model is verified against an independent model designated as the "truth" (e.g., Arnold & Dey, 1986; Atlas et al., 1985). Modern OSSEs contain the following elements: (1) a long forecast considered to be the truth or nature run (NR) that statistically simulates the real atmosphere, (2) observations simulated from the NR, and (3) a different data assimilation/forecast system that will ingest the simulated observations (Hoffman & Atlas, 2016; Masutani et al., 2010). OSSEs have investigated the impacts of many types of observing systems on weather forecasts, including wind lidar (Atlas, Hoffman, et al., 2015; Ma et al., 2015), rawinsonde data (Privé, Errico, and Tai, 2014), the data assimilation system (Kleist & Ide, 2015a, 2015b), and dropsondes for tropical storms (Atlas, Bucci, et al., 2015; Privé, Xie, et al., 2014; Qin & Mu, 2014). OSSEs investigating targeted observations generally have similar conclusions and limitations as OSEs using targeted dropsondes: Targeted dropsondes often improve forecasts on average, but results are case dependent. Occasionally, forecasts are neutral or degraded, and results are difficult to conclude with statistical confidence since targeted observations are designed to improve a specific weather event at a specific location and time and the number of cases are limited. As such, the majority of published OSSEs of targeted observations explore the impacts of observations on specific meteorological events rather than conclude whether observations can or cannot improve forecasts since the number of cases are too limited to make statistical conclusions. An advantage of OSSEs over OSEs is that it is often possible to conduct a

higher number of targeted campaigns and/or "idealized" campaigns (such as sampling over a large domain) to gain scientific understanding that may be unattainable with OSEs.

Here we use OSSEs to investigate the impacts of MW and IR satellite data and dropsonde data on global-average forecasts and forecasts for three winter storms present in the NR that impact the United States. We complete two sets of evaluations: (1) a "multiple-domain" evaluation of satellite data and idealized dropsondes (dropsondes sampled over a large region of the Pacific and Arctic Oceans) across large temporal (0- to 7-day lead times) and spatial scales (multiple verification domains for each storm) and (2) a "targeted domain" evaluation of satellite data, idealized dropsondes, and targeted dropsondes across targeted temporal (2- to 3-day lead time) and spatial scales (a single verification domain for each storm). Further details of the experimental design are discussed in section 2.3.

This study is Part 2 of a two-part OSSE research project. In Part 1 (Peevey et al., 2018), OSSEs were completed to evaluate the impact of three types of dropsonde data (temperature, humidity, and wind) over three sampling domains on the forecast accuracy of the same three winter storms present in the NR. Sampling all three types of measurements over a large idealized sampling domain produced the largest forecast improvement. Winds provided the largest individual benefit, followed by temperature and then humidity. It was also found that sampling targeted observation domains provided up to a 5% reduction in forecast energy error. Results for individual lead times and verification domains varied, with some instances of targeted observations providing no benefit or even leading to forecast degradation. Investigation into the causes of individual forecast degradations were concluded to be due to challenging meteorological features such as cutoff low-pressure systems or interactions with meteorological features outside of the sampling domains. More details on other OSE and OSSE targeted observations studies and reasons for the possible differences between them are also covered in Part 1 of this project (Peevey et al., 2018).

2. Methods and Design

2.1. OSSE Framework and Experiment Setup

We use the "T511 NR" as the atmospheric truth. The T511 NR is a 13-month uninterrupted forecast with T511 horizontal resolution (~40 km) and 91 model levels produced in 2005 by the European Centre for Medium-range Weather Forecasts using their Integrated Forecast System version cy31r1 (Masutani et al., 2007). This NR has been compared to observed climatology, and the synoptic behavior and was found to be acceptable for use in an OSSE (McCarty et al., 2012; Reale et al., 2007). The forecast was initialized at 12 UTC 1 May 2005, with the operational analysis as the initial conditions, and terminated at 00 UTC 1 June 2006 with output every 3 hr (Andersson & Masutani, 2010). (The year is somewhat arbitrary since this is a Nature Run; actual sea surface temperature observations from 2006 are used, but the atmospheric circulation is freely evolving and does not represent the Earth's atmospheric circulation in 2006). For this study, we utilize the output from January and February 2006. This OSSE setup uses perfect observations for all types of input data (conventional, satellite, and dropsonde), meaning that instrument error is neglected, which will underrepresent the error and uncertainty present in the real world. The impact of using perfect observations in an OSSE simulation has been quantified (Masutani et al., 2010). An advantage of using perfect observations is that the reduced uncertainty may allow for a clearer interpretation of forecast impact with fewer forecasts.

We use the NCEP's Global Forecast System (GFS) model Q1FY15 to ingest the simulated observations and evaluate their impact and forecast accuracy. This model is coupled with the Gridpoint Statistical Interpolation data assimilation package (Kleist et al., 2009) which was operational in 2015. The Global Data Assimilation System (GDAS) is used in the 3DVar Hybrid Ensemble Kalman Filter configuration (Wang et al., 2013). Both the Gridpoint Statistical Interpolation and GFS components are configured with a T382 horizontal resolution and 64 vertical levels (T254 physics: ~ 50 km). The Ensemble Kalman Filter component is set up with a T254 horizontal resolution and 64 vertical levels.

This OSSE framework consisting of the T511 NR with the global GFS model has been previously developed and validated (Errico et al., 2013; Privé et al., 2013) and utilized for several applications including the impacts of rawinsondes (Privé, Errico, & Tai, 2014). More details on this OSSE framework are also provided in Part 1 of this project (Peevey et al., 2018).

Table 2
Experimental Design of This Study

	Satellites assimilated		
Dropsondes assimilated	Control (2015 operational GFS)	Control minus Aqua, DMSP, and NOAA 14–19 satellites	Control minus Aqua, DMSP, NOAA 14–19, and NPP satellites
None	CTL	Base	Gap
Idealized domain	Ctl_Ideal	Base_Ideal	Gap_Ideal
Sensitivity domain	Ctl_Sens	Base_Sens	Gap_Sens
Flight path domain	Ctl_Flight	Base_Flight	Gap_Flight

Note. GFS = Global Forecast System.

2.2. Ensemble Transform Sensitivity Method

We use the Ensemble Transform Sensitivity (ETS) method to identify locations for targeted dropsonde measurements in our OSSE study. The ETS method is a fast, efficient way to identify regions sensitive to error growth (Zhang et al., 2016). The ETS method is a first-order approximation of the Ensemble Transform (ET) method (Bishop & Toth, 1999). While the ET method predicts a reduction in forecast error variation by recalculating a transformation matrix in an ensemble subspace for each possible observation deployment, the ETS method calculates the transform matrix only once, resulting in a more computationally efficient method. The ETS method then calculates the gradient, or sensitivity, of the forecast error variance over the verification region to be represented in terms of the analysis error variance. The ETS method is based on a dry total-energy norm of temperature and wind error at 200, 500, and 700 hPa (see section 2.4). A comparison OSSE study found that the ETS and ET methods identified similar regions of sensitivity, but with a 60–80% reduction in computational cost (Zhang et al., 2016).

2.3. Study Design

We conduct a series of experiments to understand the impacts of a satellite gap and of dropsonde measurements on analysis and forecast error. We investigate two satellite data denial cases ("Base" and "Gap") and three dropsonde sampling domains ("Idealized," "Sensitivity," and "Flight Path"), along with control experiments for each (Table 2). The "CTL" experiment assimilates the same satellites as the 2015 operational version of the GFS model. The Gap experiment removes the Aqua, DMSP, NOAA 15, NOAA 18, NOAA 19, and Suomi-NPP (including the Advanced Technology Microwave Sounder [ATMS] and the CrIS observations) satellites from the data assimilation. This experiment represents a worst case scenario of the loss of all U.S.-based MW and IR instruments, similar to the design of Cucurull and Anthes (2015). The Base experiment is the same as Gap, but with Suomi-NPP added back in. Thus, a comparison between Base and Gap quantifies the specific impact of a gap in MW and IR data from Suomi-NPP on weather forecasts, which is relevant given the recent Suomi-NPP satellite gap that has actually occurred.

The three satellite experiments are each run with and without added dropsondes. Dropsondes are considered over three domains denoted as "Ideal," "Sens," and "Flight." The Ideal domain samples dropsondes across a large region of the Pacific and Arctic Oceans to represent a best case or idealized impact of dropsondes (Figure 1). The Ideal domain also allows sampling over many time steps to enable a broader forecast evaluation than can be accomplished in a single study of targeted observations (Sens and Flight). Sens samples dropsondes across the sensitivity domain identified using the ETS method, which includes grid boxes where the normalized sensitivity is $\geq$0.5; the region varies for each experiment and is discussed in section 3.3. Finally, Flight uses an automated flight track algorithm to place simulated dropsondes over the most important ETS regions. The Flight Path domain assumes that the simulated aircraft flies only over ocean and samples for ~ 24 hr, which is comparable to an operational Global Hawk aircraft. More details on the three dropsonde domains are provided in Peevey et al. (2018). All dropsonde experiments assimilate temperature, wind, and humidity, since in Part 1 of this project we found that assimilating all three measurements provides the largest forecast improvement (Peevey et al., 2018). Additionally, all dropsonde experiments assimilate data in all grid boxes simultaneously for all designated cycles.

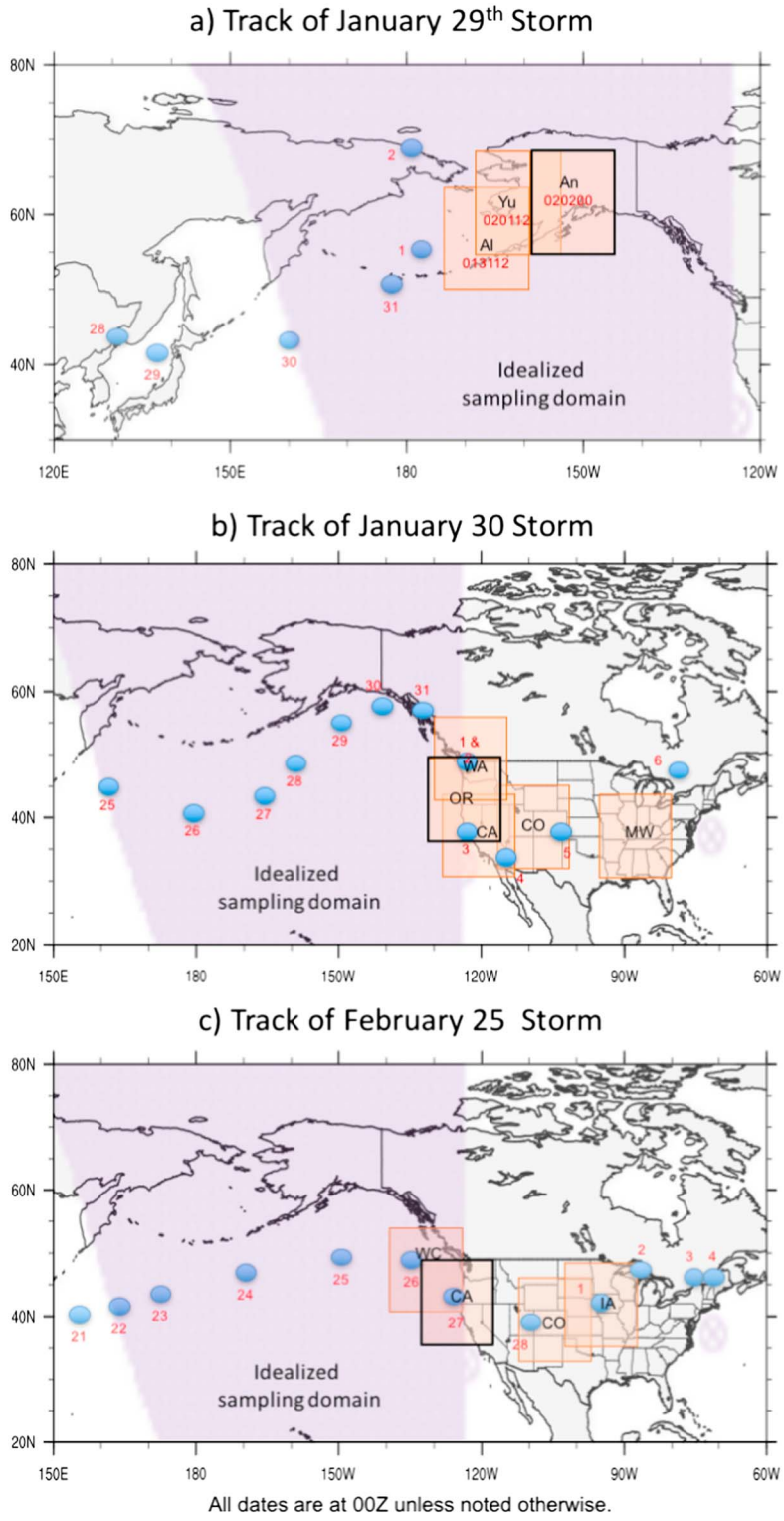

Figure 1. Schematic of storm tracks for three winter storms studied. Blue circles represent the approximate storm center at the dates below the circles. Orange boxes represent the 14° × 14° latitude-longitude verification regions studied for each storm; the black box represents the verification region used for targeted observations for each storm. (a) The 29 January verification regions are AL (Aleutians), YU (Yukon), and AN (Anchorage). (b) The 30 January verification regions are WA (Washington), OR (Oregon), CA (California), CO (Colorado), and MW (Midwest). (c) The 25 February verification regions are WC (West Coast), CA (California), CO (Colorado), and IA (Iowa). The purple region represents the Idealized dropsonde sampling domain (20–80°N).

As in Part 1 of this project (Peevey et al., 2018), we cycle all three satellite experiments (CTL, Base, and Gap) for 2 months from 1 January 2006 to 28 February 2006. Conventional observations (temperature, virtual and sensible, station pressure, humidity, and wind) from both surface and upper-level instruments are perfectly assimilated in all of the experiments. During this 2-month cycling period, there were three significant winter storms present in the NR that affected the United States: 29 January, 30 January, and 25 February (Figure 1). We evaluate each experiment's analysis and forecast accuracy during the time periods of these three storms.

We conduct two sets of evaluations: a multiple-domain evaluation and a targeted domain evaluation. The multiple-domain evaluation includes GDAS cycling for 10 days with 16 GFS forecast runs staggered 12 hr apart ranging from 0- to 7-day lead times for multiple verification regions for each storm, with the goal of understanding broad, statistical forecast impacts. The targeted domain evaluation includes GDAS cycling for 1 day with five GFS forecast runs staggered 6 hr apart ranging from 2- to 3-day lead times for a single verification region for each storm, with the goal of understanding targeted dropsonde impacts. Because sensitivity regions are strongly dependent on initialization time and verification domain, it is difficult to evaluate targeted dropsonde measurements across large spatial or temporal domains. Hence, the targeted dropsonde experiments are not included in the multiple-domain evaluation. For both sets of evaluations, we evaluate the impacts of removing satellite data and adding dropsonde data on global forecast accuracy as well as forecast accuracy of each of the three winter storms.

2.4. Metrics

The primary metric used in this study is dry total energy error (TEE; m/s; equation (1)):

$$E = \left\{ \frac{1}{2} \int_A \left[\begin{array}{l} \frac{1}{3}\left(u_{200}^2 + v_{200}^2 + \frac{c_p}{T_r} t_{200}^2 \right) + \\ \frac{1}{3}\left(u_{500}^2 + v_{500}^2 + \frac{c_p}{T_r} t_{500}^2 \right) + \\ \frac{1}{3}\left(u_{700}^2 + v_{700}^2 + \frac{c_p}{T_r} t_{700}^2 \right) \end{array} \right] \right\}^{\frac{1}{2}} \quad (1)$$

The variables u, v, and t represent the differences between the forecast and the NR for the u component of the wind, the v component of the wind, and the temperature, respectively, at three tropospheric pressure levels (200, 500, and 700 hPa). T_r is the reference temperature (300 K), c_p is the specific heat of dry air at a constant pressure (1,004 $J \cdot K^{-1} \cdot kg^{-1}$), and A is the domain area over which the calculations are averaged. TEE provides a comprehensive measure of forecast accuracy by including two state variables (wind and temperature) over three tropospheric levels. Various forms of energy error are the preferred metric for targeted observations (Gelaro et al., 2002; Hamill et al., 2013; Majumdar, 2016). The form of the energy error equation used in this study (equation (1)) is the same form as that used by the ETS method (Zhang et al., 2016), providing consistency.

We also evaluate forecast performance using other metrics: we calculate TEE using the formula of Hamill et al. (2013), which considers errors in winds and temperature at 250, 500, 750 hPa, and near the surface (10-m winds and 2-m temperature); root mean square error of 500 hPa geopotential height; sea level pressure (SLP) error; and precipitation bias (departure from mean).

3. Results

First, we investigate the impacts of removing numerous satellites from the GDAS data assimilation on analysis and forecast error during the time periods of three winter storms present in the NR. Next, we investigate the impacts of adding simulated idealized dropsondes over a large area of the Pacific/Arctic Oceans to the GDAS data assimilation to understand a "best case" (although practically unrealistic) scenario of the impact of dropsondes. Finally, we investigate targeted dropsonde observations and compare forecasts to the control experiment (CTL) and the two satellite gap experiments (Base and Gap; see Tables 1 and 2).

3.1. Impacts of a satellite gap

Analysis error maps (TEE) for each satellite experiment (CTL, Base, and Gap) and for each storm are provided in Figure 2. There is considerable analysis error over the North Pacific Ocean for all three storms, with larger

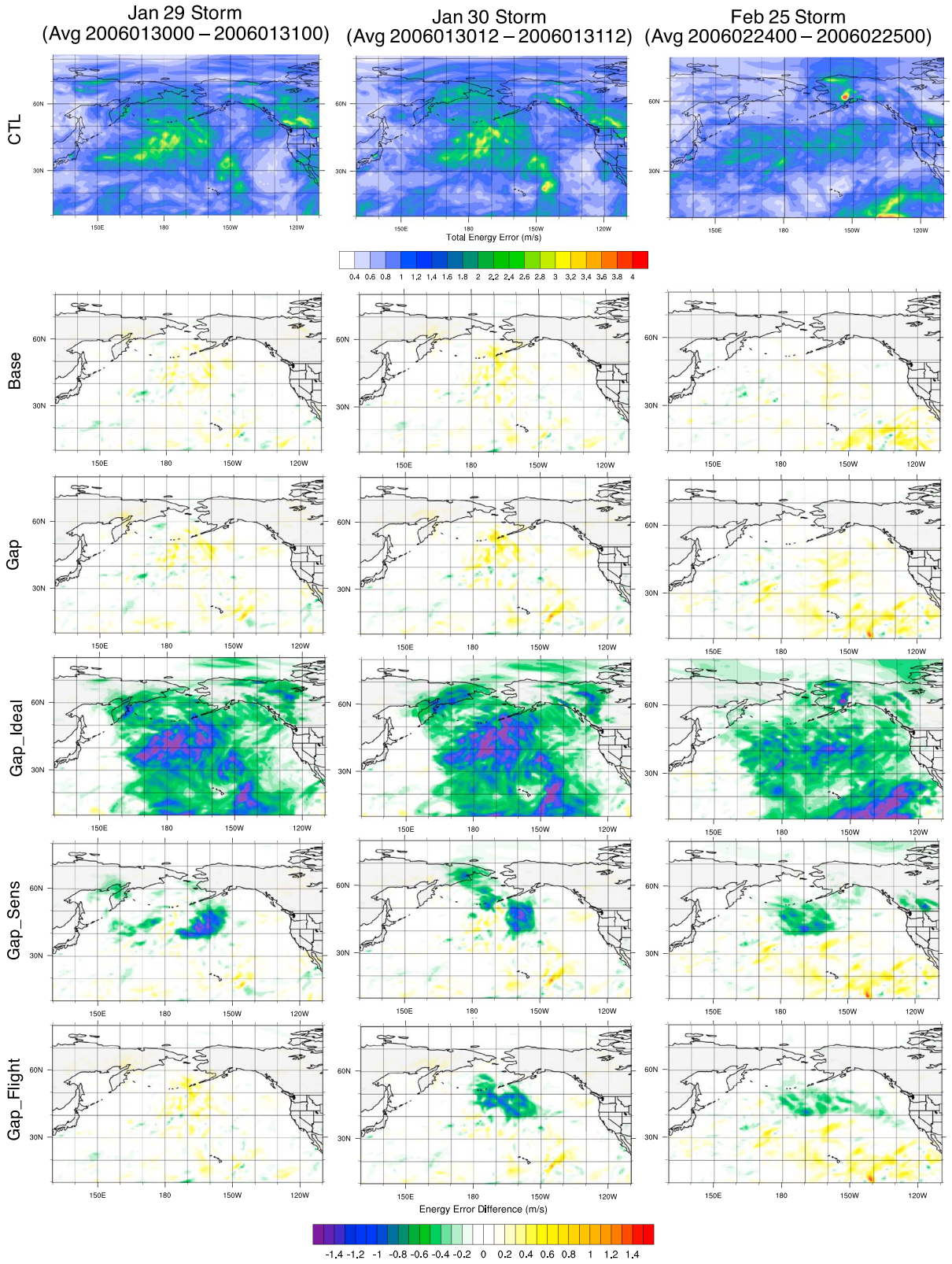

Figure 2. Global Data Assimilation System analysis maps of total energy error over the time periods of the 29 January, 30 January, and 25 February storms. Analysis error is averaged across five cycles corresponding to a 2- to 3-day lead time from the selected verification region. Maps for all experiments show absolute difference from the control, where negative indicates reduction of analysis error relative to CTL. Section 2.3 describes the three dropsonde experiments Ideal, Sens, and Flight, and Table 3 provides the number of dropsondes assimilated for each experiment.

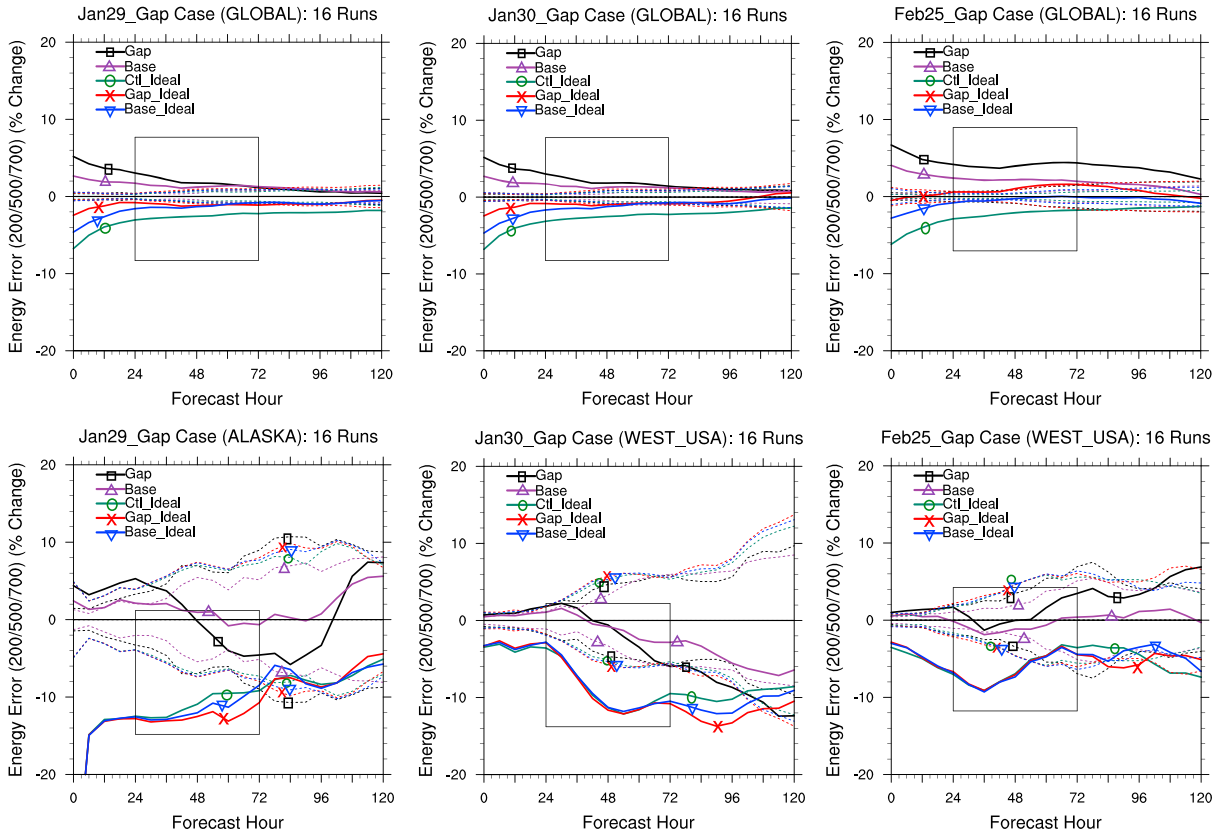

Figure 3. Percent change in total energy error for various experiments relative to CTL during each storm (columns) as a function of forecast hour, averaged globally (top panels) or averaged across a storm impact region (bottom panels) for the multiple-domain evaluation. Solid lines represent an average of 16 Global Forecast System runs staggered 12 hr apart at 0- to 7-day lead times; dotted lines represent 95% confidence interval obtained using the paired *t* test (2 times standard error from the CTL). The 29 January Storm is averaged across ALASKA (55–70°N, 195–220°E); 30 January and 25 February storms are averaged across WEST_USA (30–50°N, 235–260°E). Boxes in the top panels highlight a 1- to 3-day lead time window.

analysis error for the 29 January and 30 January storms than the 25 February storm. Analysis error over the North Pacific Ocean is a well-known issue and indeed a key reason why it has been proposed to add dropsonde measurements over this region to improve forecasts (Peevey et al., 2018). The Gap and Base experiments have slightly larger analysis error than the CTL, suggesting that removal of numerous satellites from the data assimilation may degrade forecasts. The Gap experiment has slightly larger analysis error than Base, suggesting that Suomi-NPP has a measurable impact on model analysis error.

We start with the multiple-domain evaluation of forecast accuracy averaging 16 GFS forecasts staggered 12 hr apart ranging from 0- to 7-day lead times. For all three storms, the Gap and Base experiments increase global average forecast TEE across all forecast hours (Figure 3, top panels). The Gap experiment increases global average TEE by about 6% at forecast hour 0, declining at longer forecast lead times to about a 1% increase at forecast hour 120. The increase in TEE for the Base experiment is roughly half that of the Gap experiment, suggesting that Suomi NPP is responsible for approximately 50% of the global forecast degradation from the removal of IR and MW satellites. When averaging forecast TEE over the respective domains pertaining to the locations of the storms of interest (Figure 3, bottom panels), there is much more variability, a common challenge with evaluating case-dependent targeted dropsonde evaluations. For all three winter storms, the Gap and Base experiments increase forecast TEE during the first 36 forecast hours, as expected. Immediately after the GFS forecast run has started (forecast hour 0), TEE is between 1% and 5% higher for the Gap and Base than CTL. This error grows until about forecast hour 36. However, beyond this time, there is no statistical difference in forecast TEE between the CTL, Gap, and Base experiments due to an increase in variability. Qualitatively, beyond 36 hr the Gap and Base experiments have mixed impacts on the 29 January storm, decrease TEE for the 30 January storm, and increase TEE for the 25 February storm. A comparison of Gap

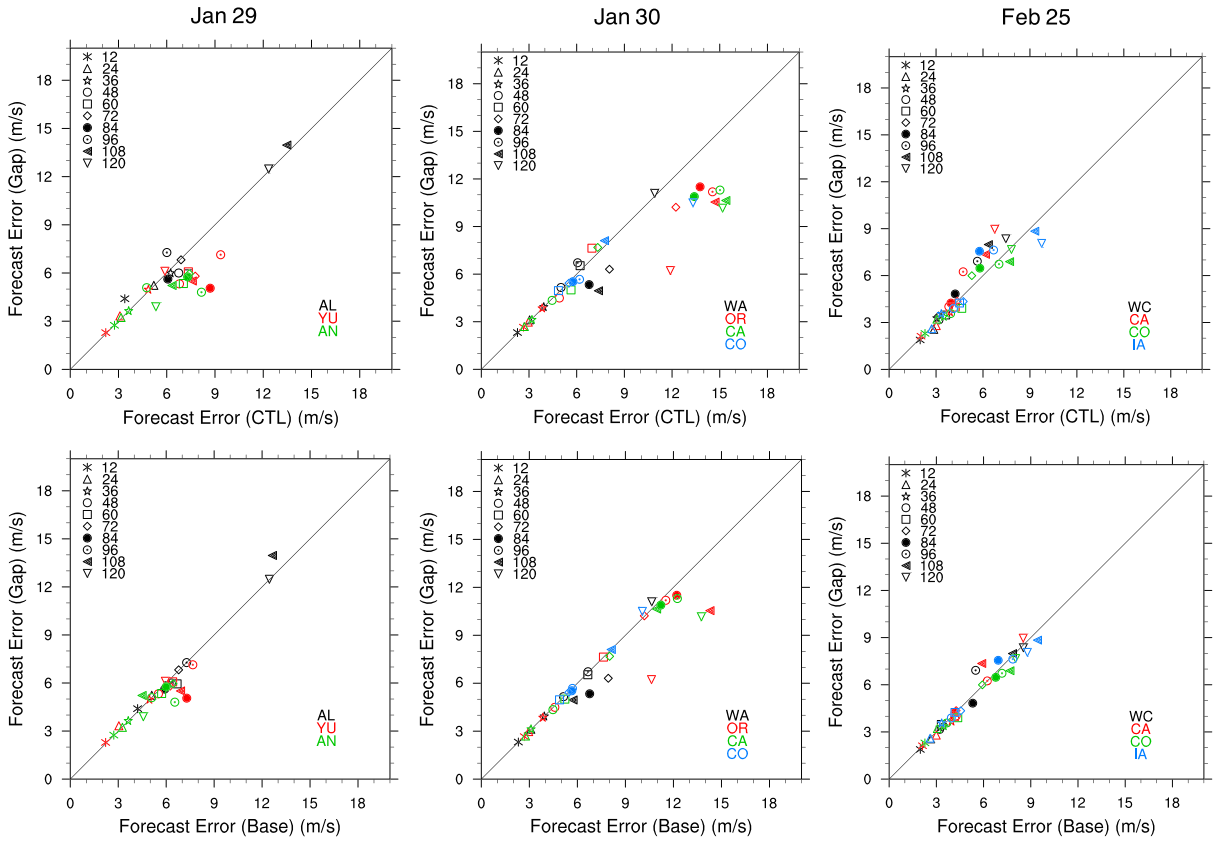

Figure 4. Scatterplots of total energy error at the designated forecast hours relative to the verification time (marker styles) at the verification regions designated in Figure 1 (marker colors) for each storm for the experiments Gap versus CTL (top panels) and Gap versus Base (bottom panels) for the multiple-domain evaluation.

to Base suggests that the Base experiment is 10–50% closer to the zero line (CTL) than Gap during the first 36 hr of the forecast lead times, suggesting that the removal of Suomi-NPP is responsible for 10–50% of the forecast degradation, while the removal of Aqua, DMSP, and NOAA 14–18 satellites is responsible for the remaining 50–90%. Beyond 36 hr, differences between Base and Gap are not statistical, but Base is always closer to the CTL than Gap.

To better understand how removal of satellites might impact forecasts of weather impacts directly related to the three winter storms, scatterplots of forecast error (TEE) for each verification region for each storm at a range of forecast lead times are provided in Figure 4. Comparing the experiments with and without Suomi NPP (Figure 4, bottom panels) suggests that removing Suomi-NPP (Gap) degrades forecasts compared to the satellite experiment with Suomi-NPP (Base) for the majority of verification regions and forecast hours for all three storms, further supporting the importance of Suomi-NPP on accurate forecasts of winter storms over the United States. There is much more variability when comparing the Gap experiment to CTL (Figure 4, top panels) with all three storms—removal of satellites (Gap) increases forecast TEE at some forecast hours and verification regions but decreases forecast TEE at others. For example, for the 29 January storm, the removal of satellites (Gap) degrades forecasts for the majority of lead times over the Aleutians verification region but actually improves forecasts over the Yukon and Anchorage verification regions. The differences between verification regions/dates are illustrated in line plots of forecast TEE as a function of verification time (Figure 5). For the 29 January storm, early in the forecast period the removal of satellites increases forecast TEE as expected. However, after about 00 UTC 30 January, forecast TEE decreases for the Gap and Base experiments until about 12 UTC 2 February, after which it increases again. Average TEE as a function of verification date has significant variability for all three storms, due in part to averaging longer lead times (up to 7 days). Even so, some of the impacts of satellite removal are statistical (95% confidence), such as the increase in forecast TEE at 00 UTC 30 January (29 January storm), the decrease in forecast TEE at 00 UTC 4 February

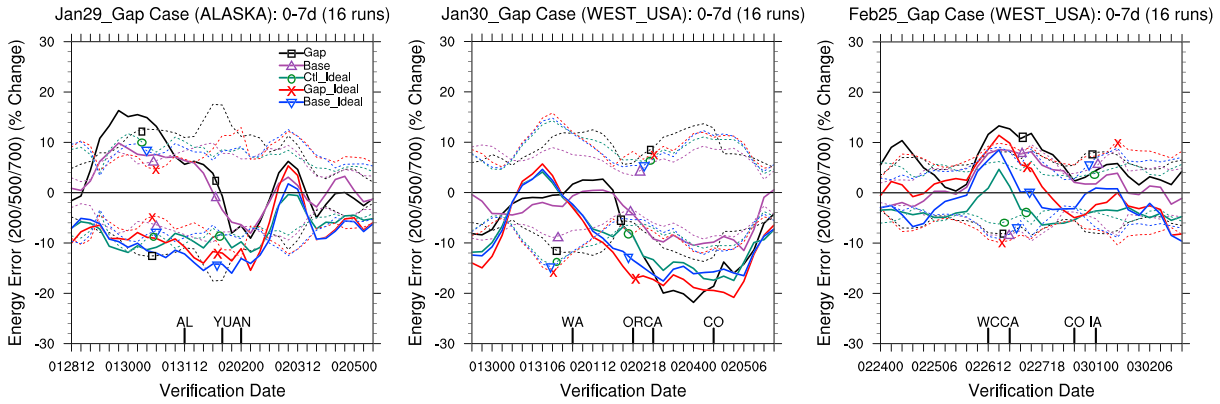

Figure 5. Same as Figure 3 bottom panels, except that *x* axis is verification date rather than forecast hour.

(30 January storm), and the increase in forecast TEE at 00 UTC 27 February (25 February storm). It is well known that the error involved with individual forecasts can vary widely and can be due to inadequate measurements, measurement error, assimilation or interpolation error, or model physics error. This suggests that it is important to evaluate models based on multiple forecast runs and case studies. Here the removal of satellites decreases forecast error for the majority of data points at the 2- to 4-day lead times for two out of three of the storms studied (29 January and 30 January). Both of these storms are initialized at nearly the same date, so the model errors or sampling noise that is occurring are likely involved with both storms. This is further discussed in section 4.

3.2. Impacts of idealized dropsondes

Here we sample dropsondes over a large portion of the Pacific/Arctic Oceans (purple domain in Figure 1) to understand the best case scenario or largest impact that dropsonde measurements might have on the forecast skill of winter storms. Again, this is not pragmatically feasible with an aircraft campaign, but this is an efficient way to estimate the possible average forecast impact of dropsondes over a large spatial/temporal domain given a limited number of cases. Analysis error maps of the Gap_Ideal experiment show a large reduction in TEE over most of the sampling region when compared to its corresponding experiment without dropsondes (Gap) as well as CTL (Figure 2). This suggests that assimilating dropsonde observations over a large idealized area significantly reduces model initialization error and can more than compensate for an increase in model analysis error when the MW and IR satellites are removed. The reduction in the analysis error translates to a reduction in the forecast error as well. This is not surprising given that dropsondes provide a direct measurement of the atmosphere as opposed to an indirect measurement from satellite radiances. Line plots also show a large reduction in global-average forecast TEE for the experiments with idealized dropsondes (Figure 3, top panels). At forecast hour 0, adding idealized dropsondes decreases global average forecast TEE by about 8% relative to its corresponding satellite experiment. For example, for the 29 January storm, Gap increases global forecast TEE by about 5% relative to the CTL, while Gap_ideal decreases global forecast TEE by about 3% relative to CTL, an improvement of 8%. The improvements fade across the forecast period, mirroring the fade in forecast degradation from satellite removal across the same period. These results suggest that dropsonde data have the potential to compensate for a loss of all U.S.-based MW and IR satellite data, although the idealized domain is unfeasibly large for a typical flight campaign. Conclusions are similar when averaging forecast TEE over the respective domains pertaining to the locations of the storms of interest (Figure 3, bottom panels): adding idealized dropsondes more than compensates for a loss of satellite data. At forecast hour 0, idealized dropsondes reduce forecast TEE by about 25% (29 January storm) and 4% (30 January and 25 February storms) relative to the CTL, and by about 30% (29 January storm) and 5% (30 January and 25 February storms) relative to their corresponding satellite gap experiments. For all three winter storms, the idealized dropsonde experiments decrease forecast TEE by ~ 10% at 1- to 3-day lead times (black boxes in Figure 3, bottom panels). The consistent improvement at these lead times is encouraging as targeted observations with flight campaigns usually obtain measurements with these similar lead times (in the next section, results are evaluated at 2- to 3-day lead times).

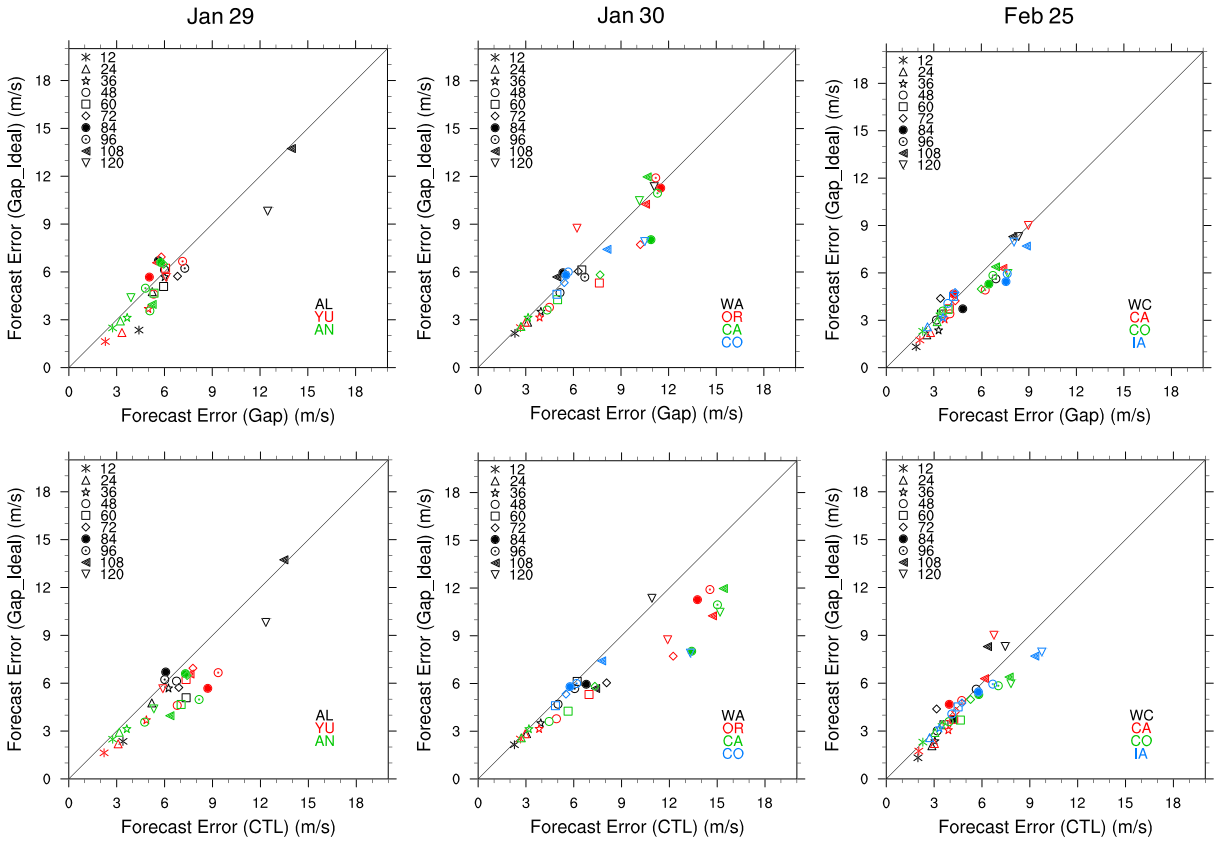

Figure 6. Same as Figure 4, except for Gap_Ideal versus Gap (top panels) and Gap_Ideal versus CTL (bottom panels).

The forecast improvements with idealized dropsonde experiments are also significant across a longer range of lead times than the satellite gap experiments: up to 72 hr (29 January storm), 96 hr (30 January storm), and 60 hr (25 February storm). All three idealized dropsonde experiments perform similarly, suggesting that adding dropsonde data of temperature, water vapor, and humidity across a large swath of the Pacific Ocean more than compensates for a loss of all U.S.-based MW and IR satellite data. Scatterplots also show that adding idealized dropsondes over a large region of the Pacific Ocean (Gap_Ideal) significantly reduces forecast error at most lead times and verification regions for all three storms. This is apparent regardless of comparing to its corresponding experiment without dropsondes (Figure 6, top panels) or comparing to the CTL (Figure 6, bottom panels).

3.3. Impacts of targeted dropsondes

Finally, we conduct a targeted domain evaluation. For each storm, we chose one verification region to conduct a detailed evaluation of the impact of targeted dropsonde sampling and satellite removal on 2- to 3-day forecast lead times: Anchorage (29 January storm), Oregon (30 January storm), and California (30 January storm; Figure 1). These are the same regions evaluated in Part 1 of this project (Peevey et al., 2018) and were chosen by meeting the following criteria: (1) strong meteorological impact (precipitation and wind), (2) large forecast error in the control run, (3) large forecast improvement in the idealized dropsonde run relative to the control run, and (4) targeted observation sensitivity maps cover a domain that is not too large to be reasonably targeted by a hypothetical 24-hr flight path. The verification regions are all 14° × 14° grid boxes, with the center of the region corresponding with the SLP minimum, to encompass the impacts of their corresponding storm. Evaluating 2- to 3-day lead times and a 24-hr flight path is somewhat representative of real flight campaigns using the Global Hawk aircraft, although in this study all of the Flight experiment dropsondes are sampled simultaneously across all five cycles. More details on the approach for assimilating dropsondes with simulated flight paths are provided by Peevey et al. (2018). Additionally, choosing these verification regions provides a further investigative opportunity: While the removal of satellite data increased forecast TEE

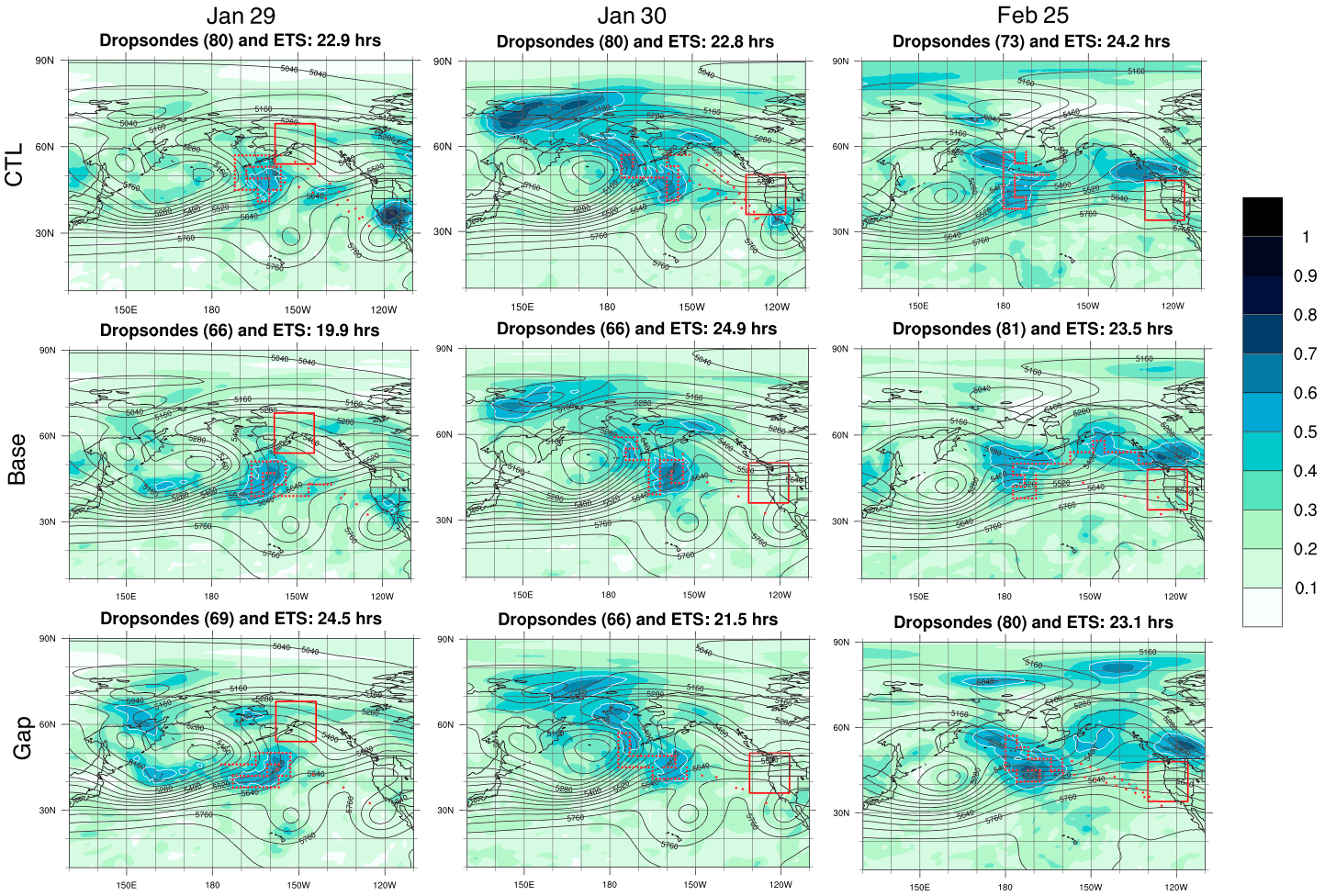

Figure 7. Normalized Ensemble Transform Sensitivity (ETS) for prespecified verification regions (red boxes) to initial perturbations, averaged over five cycles in the 2- to 3-day lead time window for each of the storms studied. Areas with sensitivity ≥0.5 are sampled in the "Sensitivity domain" experiments. Grid boxes encompassing the simulated flight paths are sampled in the "Flight Path domain" (red dots). The number of dropsondes assimilated for the flight experiments are listed in parentheses in each panel.

relative to CTL for the 25 February storm in its chosen verification region (California), it decreased forecast TEE relative to CTL for the 29 January and 30 January storms in their specified verification regions (Anchorage and Oregon, respectively). In other words, for two of the storms in this study, removal of satellites improved forecasts at the chosen verification date/region, providing an opportunity to investigate why.

The ETS method (Zhang et al., 2016) was used to identify regions sensitive to forecast error growth for the five initialization times spaced 6 hr apart corresponding to a 2- to 3-day lead time from each storm's chosen verification region. The resulting five sensitivity maps were averaged to capture the movement of the signal over the 2- to 3-day lead time and to average out spurious noise. This single normalized sensitivity map for each experiment was utilized to determine its targeted sampling domains (Figure 7). The sensitivity maps vary significantly for each storm but are quite similar across the CTL, Base, and Gap experiments. This suggests that regions sensitive to error growth may be more related to the meteorological features at the initialization time rather than changes to the data assimilated by the model. As with Part 1 of this project (Peevey et al., 2018), the Sens experiments sample all grid boxes with ETS sensitivity of ≥ 0.5 (blue and black colors in Figure 7), except for sensitivity regions over land that were east of the verification region (red box). The Flight experiments sample all grid boxes denoted by the red dots. Five GFS forecast runs were initialized from each of the dates spaced 6 hr apart, corresponding to the 2- to 3-day lead times from each verification region. The number of dropsondes assimilated in the Sens and Flight experiments varied significantly with each satellite

Table 3
Number of Dropsondes Assimilated for Each Experiment

	Number of dropsondes assimilated					
	CTL		Base		Gap	
Storm	Sensitivity region	Flight region	Sensitivity region	Flight region	Sensitivity region	Flight region
29 January	94	80	174	66	258	69
30 January	765	80	249	66	384	66
25 February	349	73	406	81	602	80

removal experiment and each storm (Table 3). In general, the sensitivity regions became larger as more satellite data were removed, probably because less data assimilated usually translates to larger forecast uncertainty, while the number of dropsondes assimilated in the Flight experiments remained relatively constant due to flight path limitations.

GDAS analysis error maps (TEE) of the three dropsonde experiments show that analysis error is reduced much more for Gap_Ideal than for Gap_Sens and Gap_Flight, as one would expect due to the much larger size of the Gap_Ideal sampling domain (Figure 2). However, there still are noteworthy reductions in analysis error for the Gap_Sens and Gap_Flight experiments relative to Gap. Analysis error for Gap_Sens and Gap_Flight are similar to or less than analysis error for CTL, suggesting that targeted dropsonde sampling may be able

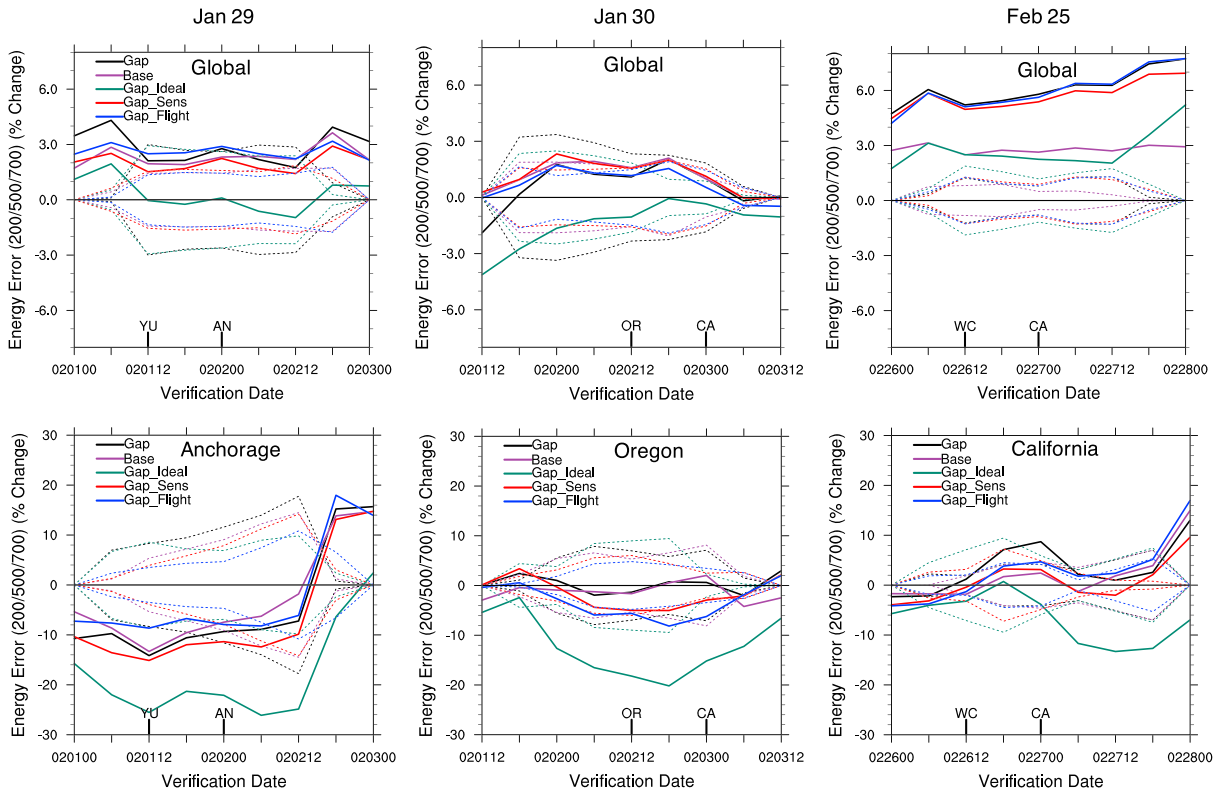

Figure 8. Percent change in total energy error relative to CTL for each storm (columns) as a function of verification date, averaged globally (top panels) or averaged across a storm impact region (bottom panels). Solid lines represent an average of five Global Forecast System runs staggered 6 hr apart at 2- to 3-day lead times; dotted lines represent 95% confidence interval obtained using the paired *t* test (2 times standard error from the CTL). The 29 January storm is averaged across ALASKA (55–70°N, 195–220°E); 30 January and 25 February storms are averaged across WEST_USA (30–50°N, 235–260°E). Designated locations in the bottom panels correspond to the verification regions in Figure 1.

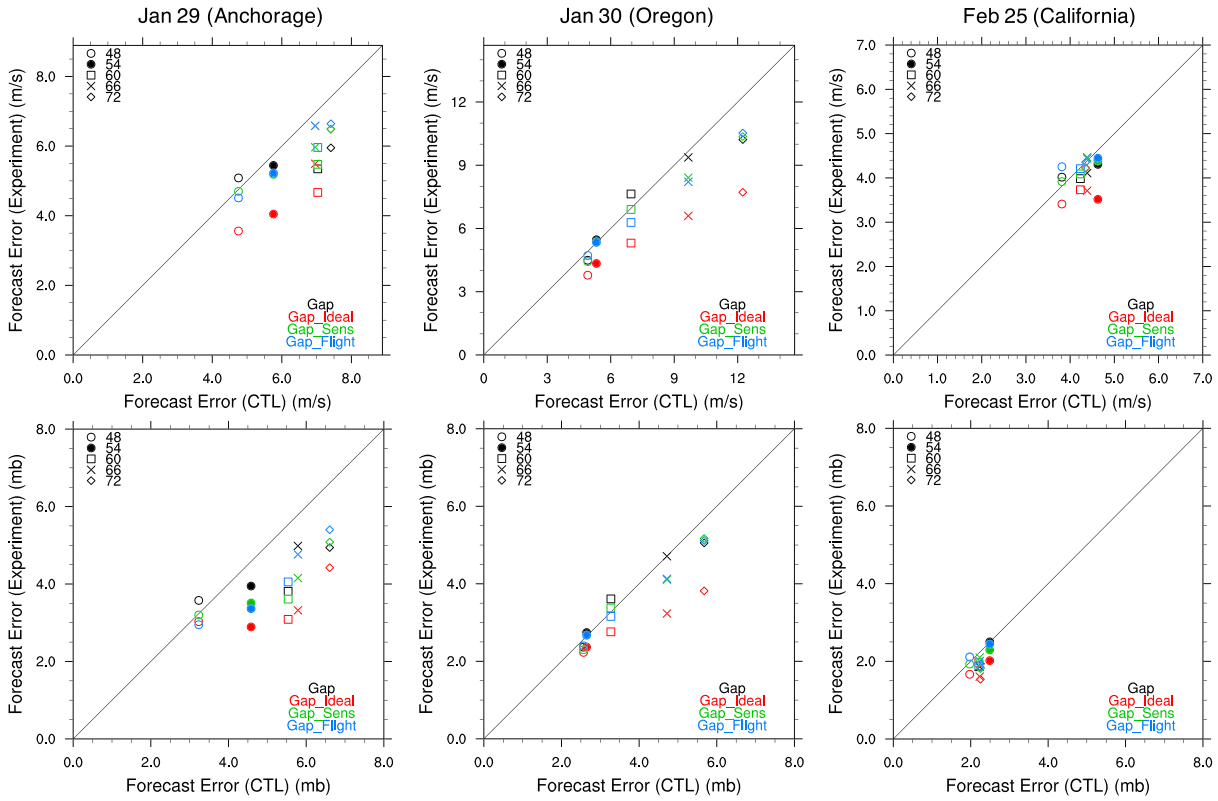

Figure 9. Scatterplots of total energy error (top panels) and sea level pressure (bottom panels) at the designated forecast hours relative to the verification time (marker styles) for each of the dropsonde experiments (marker colors) versus the CTL for each storm for the targeted dropsonde evaluation. A single verification location/time is selected for each storm (bold black boxes in Figure 1).

to compensate for a loss of all U.S.-based MW and IR satellite data. However, reductions in analysis error do not always translate to reductions in forecast error.

Regarding forecast error, for all three storms, the removal of satellites increases the global average forecast TEE by 1% to 6% at the chosen verification region (Figure 8, top panels). Suomi-NPP does not have a significant impact on forecasts for two of the storms (29 January and 30 January), while it is responsible for about 50% of the forecast degradation for the third storm (25 February). Adding idealized dropsondes significantly improves average global forecasts for all three storms, reducing TEE relative to the control by about 3% (comparing Gap to Gap_Ideal). The targeted dropsonde experiments (Gap_Sens and Gap_Flight) do not appreciably impact global average forecasts. This is expected since they are sampling a small domain relative to the globe, meaning that targeted dropsondes are probably not able to compensate for a loss of all U.S.-based MW and IR satellite data across large spatial or temporal average forecasts.

Next, we calculated TEE at the verification date corresponding to each storm's chosen verification region (Figure 8, bottom panels) and evaluated the forecast performance for each verification region. Removal of satellites (Gap and Base) *improves* forecasts for one storm (29 January), has neutral impacts on the second (30 January), and degrades forecasts for the third (25 February). Again, this highlights the challenges with evaluating case-dependent weather events and is discussed further in section 4. Despite the inherent variability, the idealized dropsonde experiments reduce forecast TEE by between 10% and 25% for all three storms, suggesting the potential benefits of dropsonde sampling under ideal conditions. The targeted dropsonde experiments (Gap_Sens and Gap_Flight) show a slight improvement in forecasts for all three storms as well: For the 29 January storm, they have roughly the same forecast performance as their corresponding satellite removal experiments with no dropsondes, which is about 10% reduced TEE than CTL. For the 30 January storm, they reduce error by about 5% relative to both the CTL and Gap, and for the 25 February storm, they reduce error by about 5% relative to Gap, which is about the same error as CTL. However, the

Table 4
Number of Model Forecasts at the Chosen Verification Locations/Times With Reduced Error Relative to the CTL or Gap Experiments for Each Storm Relative to the Total Number of Model Forecasts

	29 January		30 January		25 February		Combined	
	Versus CTL	Versus Gap	Versus CTL	Versus Gap	Versus CTL	Versus Gap	Versus CTL	Versus Gap
				ENER_ETS (TEE)				
Gap	4/5	N/A	3/5	N/A	3/5	N/A	10/15 (67%)	N/A
Gap_Ideal	5/5	4/5	5/5	5/5	5/5	5/5	15/15 (100%)	14/15 (93%)
Gap_Sens	5/5	3/5	4/5	4/5	3/5	2/5	12/15 (80%)	9/15 (60%)
Gap_Flight	5/5	3/5	4/5	3/5	2/5	1/5	11/15 (73%)	7/15 (47%)
				ENER_H				
Gap	4/5	N/A	3/5	N/A	4/5	N/A	11/15 (73%)	N/A
Gap_Ideal	5/5	4/5	5/5	5/5	5/5	5/5	15/15 (100%)	14/15 (93%)
Gap_Sens	4/5	3/5	3/5	3/5	5/5	3/5	12/15 (80%)	9/15 (60%)
Gap_Flight	5/5	3/5	4/5	3/5	3/5	1/5	12/15 (80%)	7/15 (47%)
				Z				
Gap	4/5	N/A	4/5	N/A	5/5	N/A	13/15 (87%)	N/A
Gap_Ideal	5/5	5/5	5/5	5/5	5/5	5/5	15/15 (100%)	15/15 (100%)
Gap_Sens	5/5	4/5	5/5	4/5	5/5	3/5	15/15 (100%)	11/15 (73%)
Gap_Flight	5/5	3/5	5/5	2/5	4/5	1/5	14/15 (93%)	6/15 (40%)
				SLP				
Gap	4/5	N/A	3/5	N/A	3/5	N/A	10/15 (67%)	N/A
Gap_Ideal	5/5	5/5	5/5	5/5	5/5	4/5	15/15 (100%)	14/15 (93%)
Gap_Sens	5/5	4/5	3/5	5/5	5/5	3/5	13/15 (87%)	12/15 (80%)
Gap_Flight	5/5	3/5	4/5	4/5	4/5	3/5	13/15 (87%)	10/15 (67%)

Note. TEE = total energy error; SLP = sea level pressure.

improvements are very small relative to the variability: the targeted dropsonde experiments (Gap_Sens and Gap_Flight) reduce forecast TEE relative to the CTL for the 29 January and 30 January storms (95% confidence), while there is no statistical difference for the 25 February storm for Gap_Sens and Gap_Flight relative to the Gap experiment for all three storms.

Next, we evaluate individual forecasts at 2- to 3-day lead times at each verification region for each storm using several different metrics. Scatterplots of TEE and SLP (Figure 9) show considerable variability, but most of the targeted dropsonde experiments improve forecasts relative to CTL, especially the idealized dropsonde experiments. To better quantify the performance of the different experiments, we count the number of instances that each experiment improves forecasts relative to the CTL or Gap (Table 4). Four different metrics are computed using this approach: TEE, energy error via Hamill et al. (2013), root mean square error of 500 hPa geopotential heights, and SLP. The results are quite similar regardless of the metric used. For TEE, removal of satellites (Gap) improves forecasts relative to the control for 67% of the instances (10 out of 15). Again, we would expect removal of satellites to degrade forecasts in the majority of instances, but the verification dates/locations chosen happened to coincide with a forecast improvement for two out of three storms (Figure 5). Combining results across the three storms, sampling idealized dropsondes (Gap_Ideal) improved 100% of forecasts relative to CTL and 93% of forecasts relative to Gap. Adding dropsondes over the sensitivity domain (Gap_Sens) improved 80% of forecasts relative to the CTL and 60% forecasts relative to Gap, while adding dropsondes over the simulated flight path (Gap_Flight) improved 73% of forecasts relative to the CTL and 47% of forecasts relative to Gap. The Gap_Flight results can be generally compared to what may be expected under an actual flight campaign. These results are roughly consistent with other reviews of targeting studies which find either that a small majority of forecasts are improved (Gelaro et al., 2010; Lorenc & Marriott, 2014; Majumdar, 2016) or that impacts on forecasts are generally neutral (Hamill et al., 2013). Given our limited number of case studies, and the fact that the difference between neutral results and improvements to a small majority of forecasts is so small, it is difficult to conclude whether our results support one set of studies more. In other words, we conclude that sampling dropsondes in a targeted

flight campaign provide either a neutral benefit or a small positive benefit on average, and many more cases would need to be run to conclude with more confidence on which is the case.

We also evaluated Precipitation Bias, and the differences between experiments were not statistically different from one another (not shown). Precipitation is notoriously difficult to predict and involves numerous microphysical and macrophysical processes. It is possible that the precipitation parameters are tuned to the existing model configuration, and improvements may be possible if the parameters are retuned to accommodate additional measurements in the data assimilation system. It is also possible that more cases need to be analyzed to obtain better statistical significance or that dropsonde measurements are able to improve winds and temperature but not precipitation. Additionally, the current implementation of the ETS method is designed to improve temperature and winds but not precipitation (Zhang et al., 2016).

4. Discussion and Conclusions

We conducted OSSEs to understand the impacts of removing U.S.-based MW and IR satellite data and adding dropsonde data over the Pacific/Arctic Oceans on weather forecasts. We evaluated combinations of three MW/IR satellite removal experiments (CTL, Base, and Gap) and four dropsonde experiments (CTL/no dropsondes, Ideal, Sens, and Flight; Table 2). We conducted a multiple-domain evaluation (16 GFS runs staggered 12 hr apart across 0- to 7-day lead times over three to five domains) and a targeted domain evaluation (five GFS runs staggered 6 hr apart across 2- to 3-day lead times over a single domain) for each experiment over time periods in which three winter storms present in a realistic NR impacted the United States. Our key findings are listed below along with further discussion of key findings (4) and (5) in the following paragraphs.

1. Removing all U.S.-based MW and IR satellite data increased analysis error and global forecast error across all forecast lead times by 1% to 6% and storm-specific forecast error across 0 to 36 hr lead times by 1% to 5% for all three winter storms. This is consistent with other studies that have found MW and IR satellite data to have a small positive impact to weather forecasts globally (Baker et al., 2005; Cucurull & Anthes, 2015; Lord et al., 2016; McNally, 2012). Removing Suomi-NPP in addition to numerous other MW and IR satellites (Gap) followed a similar pattern as the experiment with Suomi-NPP (Base) but with a larger magnitude: (1) when the Base experiment degraded forecasts, the Gap experiment degraded forecasts further, and (2) when the Base experiment improved forecasts, the Gap experiment improved forecasts further. Overall, data from Suomi-NPP were responsible for roughly one third of the removed satellite impacts. However, more cases are needed to conclude whether the Gap experiment should consistently be expected to have a larger impact on forecasts than Base.
2. Sampling idealized dropsonde data over a large region of the Pacific/Arctic Oceans in the multiple-domain evaluation significantly reduces analysis error, global forecast error across all lead times, and storm-specific forecast error for all three winter storms. The impacts are roughly twice as large as those in the satellite removal experiments, suggesting that an idealized sampling of many dropsonde measurements over this region has the potential to mitigate a loss of forecast accuracy from a satellite gap. This is, however, an idealized experiment and is not feasible with a typical flight campaign.
3. Assimilating targeted dropsonde measurements using the ETS method to improve forecasts at a specific verification location/time in the targeted domain evaluation showed promising results: The experiments with satellites removed and dropsondes added either improved forecasts relative to their no-dropsonde experiments (30 January and 25 February) and/or improved forecasts relative to the CTL (29 January and 30 January). Evaluating specific forecasts at 15 verification dates/locations for the three storms, we find that satellite removal degrades roughly 30% of forecasts relative to the control, while targeted dropsonde flight paths improve roughly 80% of forecasts relative to the control and 50% of forecasts relative to its corresponding run without dropsondes.
4. Across the verification period, the removal of MW and IR satellites degraded storm-specific forecasts for one storm (25 February), had a neutral impact on one storm (29 January), and improved forecasts for another (30 January).
5. The impacts of any measurements (satellite, dropsonde, or other) on specific meteorological events is difficult to conclude with statistical confidence due to the many factors that impact forecasts at any given meteorological situation, initialization time, or forecast lead time, unless many cases are run, which is generally not feasible in OSE or OSSE studies of targeted dropsonde measurements, with the exception of papers which aggregate many studies.

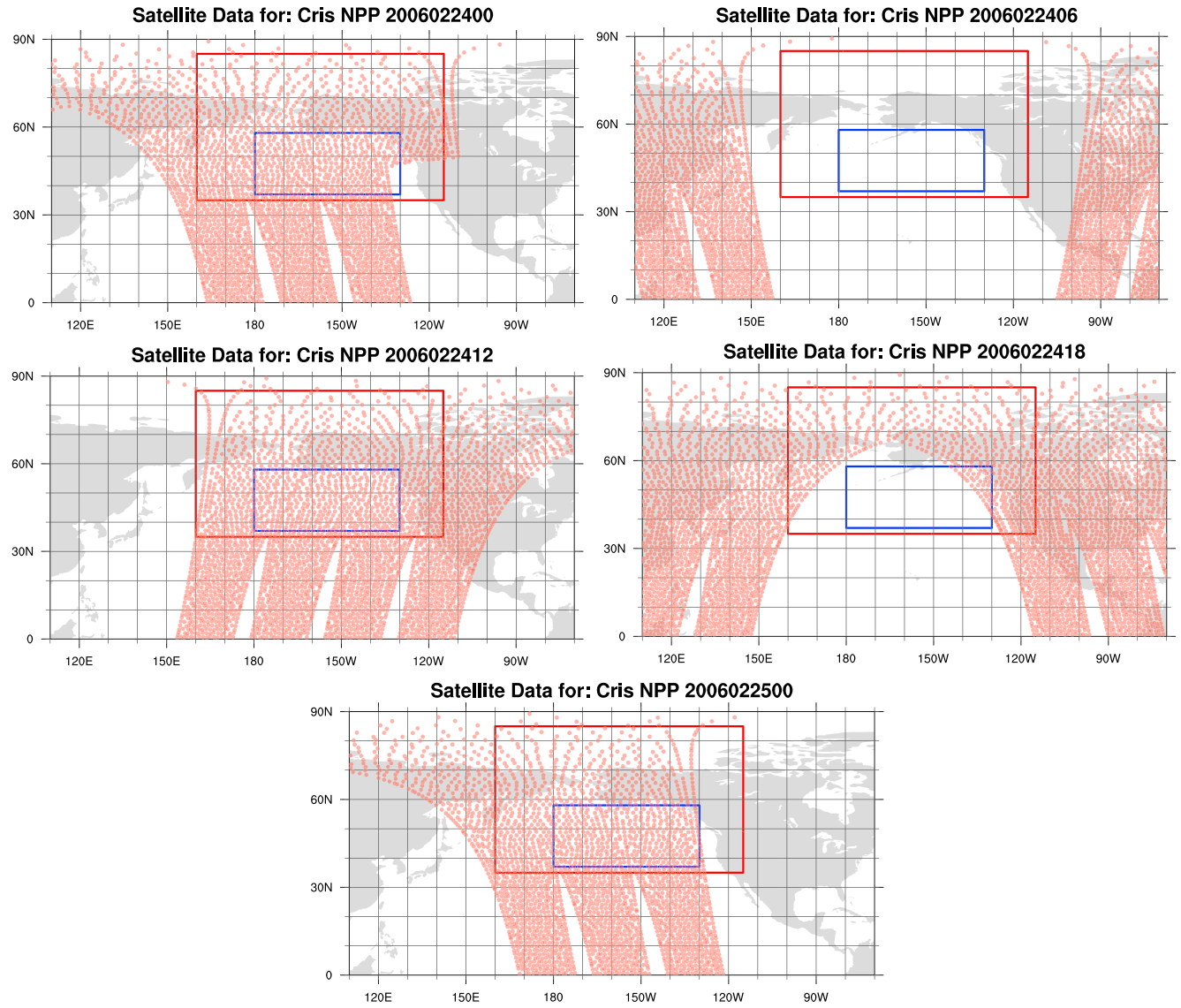

Figure 10. Satellite pass-over maps for the Cross-track Infrared Sounder (CrIS) National Polar-orbiting Partnership (NPP) instrument at each 6 hr cycling interval at the dates/times indicated, which correspond to the five cycling intervals studied in the 25 February storm. Each asterisk represents one data point. The red boxes represent an approximate range of a typical sensitivity domain, and the blue boxes represent the approximate range of a typical flight path domain.

With regards to key finding (4), a possible explanation for the large variability in forecast accuracy for the satellite removal experiments is differences in the availability of satellite data during the initialization times. Since the GDAS is cycled only for a limited time (10 days for the multiple-domain evaluation and 1 day for the targeted domain evaluation), the variability of individual forecasts may be due in part to whether the satellite(s) passed over regions important for error growth during the initialization time. For all three storms, there is considerable GDAS model analysis error over the North Pacific Ocean south of Alaska (Figure 2). For example, there are no CrIS measurements over this region for two out of five of the initialization dates relevant to the 25 February storm (Figure 10). Additionally, roughly 1% of the actual satellite data are not available for various reasons and these instances have been excluded from the assimilation in our OSSE experiment as well. If satellite data happen to be missing during the initialization times for the corresponding verification domains, this may impact results. For example, for the 30 January targeted dropsonde experiment, ATMS and CrIS data from Suomi NPP were available at 0000 and 1200 UTC but not the 0600 and 1800 UTC initialization times. The satellite pass-overs relevant to each storm were quantified using the red box as a rough

Table 5
Number of Satellite Data Points Removed Over a 24-hr Period for the Two Satellite Removal Experiments (Base and Gap)

	Satellite data removed			
	Base		Gap	
Storm	Sensitivity region (red box)	Flight region (blue box)	Sensitivity region (red box)	Flight region (blue box)
29 January	10,602	4,701	11,667	5,202
30 January	19,823	4,531	20,610	4,899
25 February	26,046	7,888	33,436	10,322

Note. This 24-hr period corresponds to a 2- to 3-day lead time from the chosen verification region. The Sensitivity region is approximated by summing all data points within the red box shown in Figure 11, while the Flight region is approximated by summing all data points with the blue box in Figure 11.

estimate of the locations important for the Sensitivity domain and the blue box as a rough estimate of the locations important for the Flight Path domain. This calculation finds significantly more satellite data points removed for the 25 February storm (Table 5). Indeed, our results show that forecasts are degraded more for the 25 February storm, consistent with the removal of more satellite data points. However, this does not explain why removal of satellites actually improves some forecasts for the 29 January and 30 January storms. Occasionally, forecasts are degraded with additional data, which may be due to observational errors, flaws in the data assimilation system, or compensating errors.

Another explanation for the large variability in forecast accuracy for the satellite removal experiments is specific meteorological events which may compensate for model initialization for forecast errors. It is clear that early in the verification period for the 29 January case (Figure 5), removal of satellites increases forecast TEE as expected. After about 00 UTC 30 January, forecast TEE decreases until about 12 UTC Feb 2, where satellite removal experiments actually improve forecasts relative to the control. This feature appears to propagate westward and impact results for the 30 January case as well, suggesting that a single meteorological feature might be causing the improvement in forecasts for the satellite removal experiments for both storms. We explore latitude-longitude contour plots at two verification times: 00 UTC 30 January (when Gap increases TEE by about 18%) and 00 UTC 2 February (when Gap decreases TEE by about 8%; Figure 11). At both verification times, TEE for the control is largest in and near three significant meteorological features: a cutoff low near Baja California, the 29 January storm of interest, and the 30 January storm of interest. TEE for the Gap case is higher than CTL at many places as expected but is lower at two noteworthy locations/times: the cutoff low near Baja California at 00 UTC 30 January and the 30 January storm (the Pacific Ocean near the Pacific Northwest). It appears that removing satellite radiance data improves the forecasts for the cutoff low, which in turn improves forecasts for the 30 January storm. The cause is unclear, but cutoff lows are difficult to model accurately (see Peevey et al., 2018). It is also possible that removal of satellite data introduces compensating errors that improves forecasts. Either way, these results highlight the variability inherent with researching case studies of specific weather events.

With regards to key finding (5), our results varied significantly with storm case, forecast lead time, and initialization date. In Part 1 of this research project, investigation into the causes of individual forecast degradations were concluded to be due to challenging meteorological features such as cutoff lows or interactions with meteorological features outside of the sampling domains (Peevey et al., 2018). This variability seems unavoidable in the field of weather forecasting and is why many storm cases, forecast lead times, and initialization dates need to be considered to provide robust statistical results. This robustness is difficult to achieve with targeted dropsonde studies because they are by design targeting specific meteorological events. Hence, it should not be surprising that many published studies of targeting specific weather events either do not calculate statistical confidence or do not find statistically significant differences, and there continues to be debate regarding the value of targeted observations. Our results are roughly consistent with other reviews of targeting studies which find either that a small majority of forecasts are improved (Gelaro et al., 2010; Lorenc & Marriott, 2014; Majumdar, 2016) or that impacts on forecasts are generally neutral (Hamill et al., 2013). If there indeed is a meaningful discrepancy across studies, it may be due to

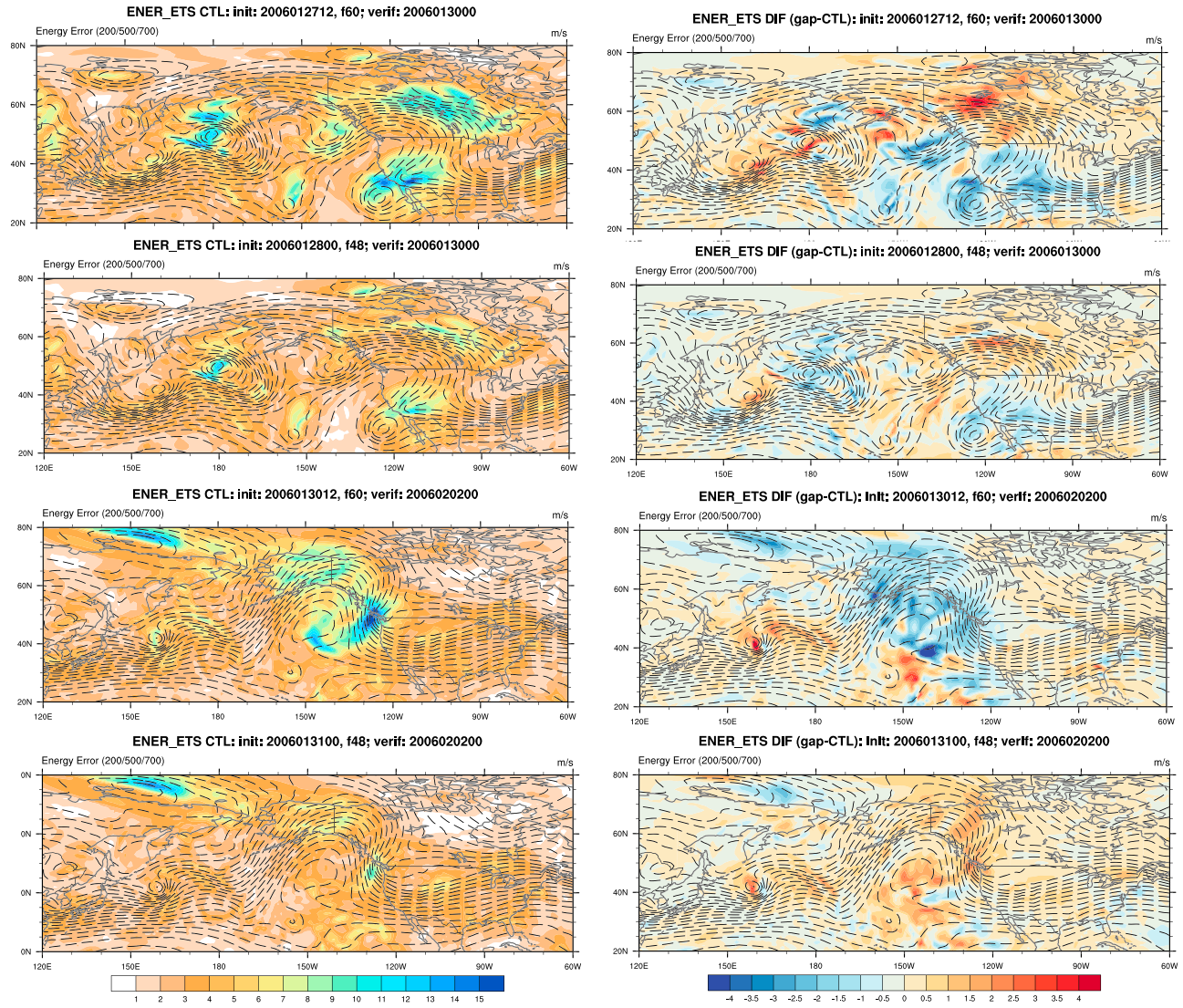

Figure 11. Latitude-longitude contour plots of the CTL total energy error (left column) and the difference between Gap and the CTL (Gap − CTL; right column) for the designated initialization times and verification times.

differences in NWP models or data assimilation, verification or targeting techniques, assumption of perfect observations in OSSEs, or meteorological features of the storms studied. For example, Hamill et al. (2013) utilized a relatively advanced model (European Centre for Medium-range Weather Forecasts with 4DVar) but only assimilated 776 dropsondes (an average of eight dropsondes per flight), whereas in our OSSE study we utilized the Global GFS with 3DVar but assimilated about 70 dropsondes per flight owing to the longer flight paths attainable using Global Hawk aircraft. Majumdar (2016) argued that advances in NWP, limitations in the range of targeted observations, the flow regime for the storm of interest, and model errors may all contribute to inconclusive or disagreeing results of targeted observations. Perceived disagreements between studies may also be due to low statistical confidence.

Our results highlight that many cases need to be studied before concluding with statistical confidence what impacts, if any, measurements (satellite, dropsonde, or other) have on forecasts of specific weather events. This also means that decisions on whether to conduct field campaigns to gather targeted observations should recognize the possibility of neutral or degraded forecasts for individual storms even if the targeted observations should improve forecasts on average. More work needs to be conducted to understand the circumstances which lead to targeted observations to be more (or less) useful to help inform the decision

process. Likewise, more work needs to be conducted to understand whether redundant satellite observations are valuable to forecast accuracy of individual storm events over the United States. Finally, future work should investigate the applicability of more advanced statistical methods to better quantify whether targeted observations—as well as satellite data—impact specific weather events, given a limited number of case studies that can be completed.

Acknowledgments
We gratefully acknowledge support for this work provided by NOAA under the auspices of the Cooperative Agreements NA14OAR4320125 for CIRA and NA15OAR4320137 for CIRES. Model output used here are retained in the NOAA/NCEP mass store for a minimum of 5 years and available by contacting the lead author of this work. Simulated conventional and satellite data were produced by NCEP/EMC and JCSDA. The authors thank Sean Casey for providing the simulated conventional and satellite data, Kate Howard at the National Centers for Environmental Prediction (NCEP) for her assistance in setting up the NCEP GFS Model, Lidia Cucurull for providing input on the initial experimental design, and Bob Atlas for recommending evaluation metrics. The authors thank Betsy Weatherhead, Zoltan Toth, and anonymous reviewers for their thoughtful suggestions to improve this manuscript.

References

Aberson, S. D. (2008). Large forecast degradations due to synoptic surveillance during the 2004 and 2005 hurricane seasons. *Monthly Weather Review*, *136*, 3138–3150.

Aberson, S. D. (2010). Ten years of hurricane synoptic surveillance (1997–2006). *Monthly Weather Review*, *138*, 1536–1549.

Andersson, E., & Masutani, M. (2010). Collaboration on Observing System Simulation Experiments (joint OSSE). *Bulletin of the American Meteorological Society*, *67*(6), 687–695.

Arnold, C. J., & Dey, C. (1986). Observing-System Simulation Experiments: Past, present, and future. *Bulletin of the American Meteorological Society*, *67*, 687–695.

Atlas, R., Bucci, L., Annane, B., Hoffman, R., & Murillo, S. (2015). Observing System Simulation Experiments to assess the potential impact of new observing systems on hurricane forecasting. *Marine Technology Society Journal*, *49*, 140–148. https://doi.org/10.4031/MTSJ.49.6.3

Atlas, R., Hoffman, R. N., Ma, Z., Emmitt, G. D., Wood, S. A. Jr., Greco, S., et al. (2015). Observing System Simulation Experiments (OSSEs) to evaluate the potential impact of an optical autocovariance wind lidar (OAWL) on numerical weather prediction. *Journal of Atmospheric and Oceanic Technology*, *32*, 1593–1613. https://doi.org/10.1175/JTECH-D-15-0038.1

Atlas, R., Kalnay, E., Baker, W. E., Susskind, J., Reuter, D., & Halem, M. (1985). Simulation studies of the impact of future observing systems on weather prediction. In *Preprints, Seventh Conf. on Numerical Weather Prediction, Montreal, QC, Canada American Meteorological Society* (pp. 145–151).

Baker, N. L., Hogan, T. F., Campbell, W. F., Pauley, R. L., & Swadley, S. D. (2005). The impact of AMSU-A radiance assimilation in the U.S. Navy's Operational Global Atmospheric Prediction System (NOGAPS) (Naval Research Laboratory Rep. NRL/MR/7530–05-8836, 22 pp.).

Bishop, C. H., & Toth, Z. (1999). Ensemble transformation and adaptive observations. *Journal of the Atmospheric Sciences*, *56*, 1748–1765.

Boukabara, S.-A., Garrett, K., & Kumar, V. K. (2016). Potential gaps in the satellite observing system coverage: Assessment of impact on NOAA's numerical weather prediction overall skills. *Monthly Weather Review*, *144*(7), 2547–2563. https://doi.org/10.1175/MWR-D-16-0013.1

Cucurull, L., & Anthes, R. A. (2015). Impact of loss of U.S. microwave and radio occultation observations in operational numerical weather prediction in support of the U.S. data gap mitigation activities. *Weather Forecasting*, *30*, 255–269. https://doi.org/10.1175/WAF-D-14-00077.1

Errico, R., Yang, R., Prive, N., Tai, K.-T., Todling, R., Sienkiewicz, M., & Guo, J. (2013). Development and validation of Observing-System Simulation Experiments at NASA's Global Modeling and Assimilation Office. *Quarterly Journal of the Royal Meteorological Society*, *139*(674), 1162–1178. https://doi.org/10.1002/qj.2027

Garrett, K. (2013). *Forecast impact assessments of SNPP ATMS*, *JCSCA Quarterly, No. 43* (pp. 3–4). College Park, MD: Joint Center for Satellite Data Assimilation.

Gelaro, R., Langland, R. H., Pellerin, S., & Todling, R. (2010). The THORPEX Observation Impact Intercomparison Experiment. *Monthly Weather Review*, *138*, 4009–4025. https://doi.org/10.1175/2010MWR3393.1

Gelaro, R., Rosmond, T. E., & Daley, R. (2002). Singular vector calculations with an analysis error variance metric. *Monthly Weather Review*, *130*, 1166–1186.

Hamill, T., Yang, F., Cardinali, C., & Majumdar, S. (2013). Impact of targeted Winter Storm Reconnaissance dropwindsonde data on midlatitude numerical weather predictions. *Monthly Weather Review*, *141*(6), 2058–2065. https://doi.org/10.1175/MWR-D-12-00309.1

Hoffman, R., & Atlas, R. (2016). Future Observing System Simulation Experiments. *Bulletin of the American Meteorological Society*, *97*(9), 1601–1616. https://doi.org/10.1175/BAMS-D-15-00200.1

Irvine, E., Gray, S., Methven, J., & Renfrew, I. (2011). Forecast impact of targeted observations: sensitivity to observation error and proximity to steep orography. *Monthly Weather Review*, *139*(1), 69–78. https://doi.org/10.1175/2010MWR3459.1

Irvine, E. A., Gray, S. L., Methven, J., Renfrew, I. A., Bovis, K., & Swinbank, R. (2009). The impact of targeted observations made during the Greenland Flow Distortion Experiment. *Quarterly Journal of the Royal Meteorological Society*, *135*, 2012–2029.

Kleist, D., Parrish, D., Derber, J., Treadon, R., Wu, W.-S., & Lord, S. (2009). Introduction of the GSI into the NCEP global data assimilation system. *Weather Forecasting*, *24*, 1691–1705.

Kleist, D. T., & Ide, K. (2015a). An OSSE-based evaluation of hybrid variational-ensemble data assimilation for the NCEP GFS. Part I: System description and 3D-hybrid results. *Monthly Weather Review*. https://doi.org/10.1175/MWR-D-13-00351.1

Kleist, D. T., & Ide, K. (2015b). An OSSE-based evaluation of hybrid variational-ensemble data assimilation for the NCEP GFS. Part II: 4DEnVar and hybrid variants. *Monthly Weather Review*. https://doi.org/10.1175/MWR-D-13-00350.1

Langland, R., Gelaro, R., Rohaly, G., & Shapiro, M. (1999). Targeted observations in FASTEX: Adjoint-based targeting procedures and data impact experiments in IOP17 and IOP18. *Quarterly Journal of the Royal Meteorological Society*, *125*, 3241–3270. https://doi.org/10.1002/qj.49712556107

Langland, R., Toth, Z., Gelaro, R., Szunyogh, I., Shapiro, M. A., Majumdar, S. J., et al. (1999). The North Pacific Experiment (NORPEX-98): Targeted observations for improved North American weather forecasts. *Bulletin of the American Meteorological Society*, *80*(7), 1363–1384. https://doi.org/10.1175/1520-0477(1999)080%E2%9F%A81363:TNPENT%E2%9F%A92.0.CO;2

Lord, S., Gayno, G., & Zhang, F. (2016). Analysis of an observing system experiment for the Joint Polar Satellite System. *Bulletin of the American Meteorological Society*. https://doi.org/10.1175/BAMS-D-14-00207.1

Lorenc, A. C., & Marriott, R. T. (2014). Forecast sensitivity to observations in the Met Office global numerical weather prediction system. *Quarterly Journal of the Royal Meteorological Society*, *140*, 209–224. https://doi.org/10.1002/qj.2122

Ma, Z., Riishojgaard, L. P., Masutani, M., Woollen, J. S., & Emmitt, G. D. (2015). Impact of different satellite wind lidar telescope configurations on NCEP GFS forecast skill in Observing System Simulation Experiments. *Journal of Atmospheric and Oceanic Technology*, *32*, 478–495. https://doi.org/10.1175/JTECH-D-14-00057.1

Majumdar, S. (2016). A review of targeted observations. *Bulletin of the American Meteorological Society*, 2287–2303. https://doi.org/10.1175/BAMS-D-14-00259.1

Masutani, M., Andersson, E., Terry, J., Reale, O., Jusem, J. C., Riishojgaard, L. P., et al. (2007). Progress in Observing Systems Simulation Experiments (a new nature run and international collaboration). In *22nd Conf. on Weather Analysis and Forecasting/18th Conf. on Numerical Weather Prediction, Park City, UT, Amer. Meteor. Soc., 12B.5*. Retrieved from https://ams.confex.com/ams/22WAF18NWP/webprogram/Paper124080.html

Masutani, M., Woollen, J., Lord, S., Emmitt, G., Kleespies, T., Wood, S. A., et al. (2010). Observing System Simulation Experiments at the National Centers for Environmental Prediction. *Journal of Geophysical Research*, *115*, D07101. https://doi.org/10.1029/2009JD012528

McCarty, W., Errico, R., & Gelaro, R. (2012). Cloud coverage in the joint OSSE nature run. *Monthly Weather Review*, *140*, 1863–1871. https://doi.org/10.1175/MWR-D-11-00131.1

McNally, A. P. (2012). Observing system experiments to assess the impact of possible future degradation of the Global Satellite Observing Network (ECMWF Tech. Memo. 672, 20 pp.). Retrieved from www.ecmwf.int/sites/default/files/elibrary/2012/11085-observing-system-experiments-assess-impact-possible-future-degradation-global-satellite.pdf

Peevey, T. R., English, J. M., Cucurull, L., Wang, H., & Kren, A. C. (2018). Improving winter storm forecasts with Observing System Simulation Experiments (OSSEs): Part 1, An idealized case study of three US storms. *Monthly Weather Review*. https://doi.org/10.1175/MWRD-17-0160.1, in press

Privé, N., Xie, Y., Woollen, J., Koch, S., Atlas, R., & Hood, R. (2013). Evaluation of the Earth Systems Research Laboratory's Observing System Simulation Experiment system. *Tellus*, *65A*, 19011. https://doi.org/10.3402/tellusa.v65i0.19011

Privé, N. C., Errico, R., & Tai, K.-S. (2014). The impact of increased frequency of rawinsonde observations on forecast skill investigated with an Observing System Simulation Experiment. *Monthly Weather Review*, *142*, 1823–1834. https://doi.org/10.1175/MWR-D-13-00237.1

Privé, N. C., Xie, Y., Koch, S., Atlas, R., Majumdar, S. J., & Hoffman, R. (2014). An Observing System Simulation Experiment for the unmanned aircraft system data impact on tropical cyclone track forecasts. *Monthly Weather Review*, *142*, 4357–4363. https://doi.org/10.1175/MWR-D-14-00197.1

Qin, X. H., & Mu, M. (2014). Can adaptive observations improve tropical cyclone intensity forecasts? *Advances in Atmospheric Sciences*, *31*(2), 252–262. https://doi.org/10.1007/s00376-013-3008-0

Reale, O., Terry, J., Masutani, M., Andersson, E., Riishojgaard, L., & Jusem, J. (2007). Preliminary evaluation of the European Centre for Medium-range Weather Forecasts (ECMWF) nature run over the tropical Atlantic and African monsoon region. *Geophysical Research Letters*, *34*, L22810. https://doi.org/10.1029/2007GL031640

Simmons, A. J., & Hollingsworth, A. (2002). Some aspects of the improvement in skill of numerical weather prediction. *Quarterly Journal of the Royal Meteorological Society*, *128*, 647–677.

Szunyogh, I., Toth, Z., Morss, R., Majumdar, S., Etherton, B., & Bishop, C. (2000). The effect of targeted dropsonde observations during the 1999 Winter Storm Reconnaissance program. *Monthly Weather Review*, *128*(10), 3520–3537. https://doi.org/10.1175/1520-0493(2000)128<3520:TEOTDO>2.0.CO;2

Szunyogh, I., Toth, Z., Zimin, A., Majumdar, S., & Persson, A. (2002). Propagation of the effect of targeted observations: The 2000 Winter Storm Reconnaissance program. *Monthly Weather Review*, *130*(5), 1144–1165. https://doi.org/10.1175/1520-0493(2002)130h1144:POTEOTi2.0.CO;2

Wang, X., Parrish, D., Kleist, D., & Whitaker, J. (2013). GSI 3D Var based ensemble-variational hybrid data assimilation for NCEP Global Forecast System: Single-resolution experiments. *Monthly Weather Review*, *141*, 4098–4117. https://doi.org/10.1175/MWR-D-12-00141.1

Zhang, Y., Xie, Y., Wang, H., Chen, D., & Toth, Z. (2016). Ensemble transform sensitivity method for adaptive observations. *Advances in Atmospheric Sciences*, *33*(1), 10–20. https://doi.org/10.1007/s00376-015-5031-9

Zou, X., Weng, F., Zhang, B., Lin, L., Qin, Z., & Tallapragada, V. (2013). Impacts of assimilation of ATMS data in HWRF on track and intensity forecasts of 2012 four landfall hurricanes. *Journal of Geophysical Research: Atmospheres*, *118*, 11,558–11,576. https://doi.org/10.1002/2013JD020405